Linking Approximation Theory and Statistical Estimation in Wavelet Image Processing

E. Le Pennec (Université Paris Diderot)

Universidad de La Habana / January 2008

1 Orthonormal Basis and Approximation Theory

1.1 Decomposition in a basis

Decomposition in a basis: $(b_j)_{j\in\mathcal{J}}$ basis of $L^2([0,1])$.

$\forall f \in L^2([0,1]$,

$$f = \sum_{j\in\mathcal{J}} \langle f, b_j\rangle b_j$$

with $\langle f, b_j\rangle = \int_0^1 f(t)\overline{b_j(t)}dt$. The Parseval-Bessel norm energy conservation equality yields

$$\|f\|^2 = \sum_{j\in\mathcal{J}} |\langle f, b_j\rangle|^2$$

1.2 Linear and non linear approximation

For any $I \in \mathcal{P}(\mathcal{J})$, one defines $f_{|I}$ by

$$f_{|I} = \sum_{j\in I} \langle f, b_j\rangle b_j \quad .$$

The question is now how to select "a good subset" of a given size M.

A first possibility is to select the N first coefficients. This assume that there is a natural order on the basis elements and thus $\mathcal{J} \equiv \mathcal{N}$. The resulting function f_M^L is thus

$$f_M^L = \sum_{j=0}^{N-1} \langle f, b_j\rangle b_j$$

where the L in the notation is here to stress that it is a linear approximation.

A second possibility is to try to minimize the error between f and $f_{|I}$. Indeed by the Parseval-Bessel equality, we know that

$$\|f - f_{|I}\|^2 = \sum_{j\notin I} |\langle f, b_j\rangle|^2 = \|f\|^2 - \sum_{j\in I} |\langle f, b_j\rangle|^2 \quad .$$

To minimize this error, it appears thus clearly that one should put in the best subset I_0 the M largest coefficients. If we denote by $T(M)$ the absolute value of the Mth largest coefficients (and we suppose that all the coefficients are differents) then the best subset I_0 is defined by

$$I_0 = \{j, |\langle f, b_j \rangle| \geq T(M)\} \quad .$$

Conversely for any T, that is called a threshold, the set

$$I_0 = \{j, |\langle f, b_j \rangle| \geq T\} \quad .$$

is the solution that minimize $\|f - f_{|I}\|^2$ for a certain constraint $|I| \leq M(T)$ where $|I|$ denotes the cardinal of I. Indeed I_0 is the minimizer in I of the corresponding Lagrangian form

$$\mathcal{L}(I, T) = \|f - f_{|I}\|^2 + T^2 |I| \quad .$$

From now on, we will denote f_M^{NL} the best non linear approximation with M terms whose absolutes values are larger than $T(M)$:

$$f_M^{NL} = \sum_{|\langle f, b_j \rangle| \geq T(M)} \langle f, b_j \rangle b_j$$

and by f_T the approximation of f by the $M(T)$ coefficients larger that T:

$$f_T = \sum_{|\langle f, b_j \rangle| \geq T} \langle f, b_j \rangle b_j \quad .$$

1.3 Approximation error decay

A natural question that arises is how fast does the approximation error decays when one increases M or decrease T. This is the subject of the approximation theory and one can relates these decays to the decay of the coefficients themselves.

1.3.1 The linear case

For the linear approximation, the following Theorem holds

Theorem 1.

$$\exists C > 0, \forall M, \quad \|f - f_M^L\|^2 \leq C M^{-\beta} \Leftrightarrow \exists C > 0, \epsilon \lambda > 1, \forall M, \quad \sum_{M+1}^{\lambda M} |\langle f, b_j \rangle|^2 \leq C M^{-\beta} \quad .$$

Proof. The direct implication is obvious as

$$\|f - f_M^L\|^2 = \sum_{M+1}^{+\infty} |\langle f, b_j \rangle|^2 \geq \sum_{M+1}^{\lambda M} |\langle f, b_j \rangle|^2$$

for any λ. To obtain the reverse implication, it is sufficient to verify that

$$\|f - f_M^L\|^2 = \sum_{M+1}^{+\infty} |\langle f, b_j\rangle|^2 \leq \sum_{k=0}^{+\infty} \sum_{\lambda^k M+1}^{\lambda^{k+1} M} |\langle f, b_j\rangle|^2$$

$$\|f - f_M^L\|^2 \leq \sum_{k=0}^{+\infty} C(\lambda^k M)^{-\beta} \leq C \sum_{k=0}^{+\infty} C(\lambda^k)^{-\beta} M^{-\beta} \leq \frac{C}{1-\lambda^{-\beta}} M^{-\beta}$$

□

A sufficient condition is given by the classical Sobolev like condition

$$\sum_{j\in\mathbb{N}} j^\beta |\langle f, b_j\rangle|^2 \leq C \quad .$$

Indeed,

$$\sum_{j\in\mathbb{N}} j^\beta |\langle f, b_j\rangle|^2 \geq \sum_{j=M+1}^{+\infty} j^\beta |\langle f, b_j\rangle|^2 \geq M^\beta \sum_{j=M+1}^{+\infty} |\langle f, b_j\rangle|^2 \quad .$$

1.3.2 The non linear case

For the non linear case, the analysis is much more precise. Reusing the notations $T(M)$ for the absolute value of the Mth largest coefficients and $M(T)$ for the number of coefficients larger than T, the following Theorem holds

Theorem 2. *For any $\beta > 0$, for any $f \in L^2$ the following properties are equivalents:*

1. $\exists C, \forall M, \quad \|f - f_M^{NL}\|^2 \leq C M^{-\beta}$
2. $\exists C, \forall T > 0, \quad \inf_M \|f - f_M^{NL}\|^2 + T^2 M \leq C(T^2)^{\beta/(\beta+1)}$
3. $\exists C, \forall T > 0, \quad \|f - f_T\|^2 + T^2 M(T) \leq C(T^2)^{\beta/(\beta+1)}$
4. $\exists C, \forall T > 0, \quad \|f - f_T\|^2 \leq C(T^2)^{\beta/(\beta+1)}$
5. $\exists C, \forall T > 0, \quad M(T) \leq C(T^2)^{-1/(\beta+1)}$
6. $\exists C, \forall M, \quad T(M) \leq C M^{(\beta+1)/2}$

Proof. $1 \implies 2$:

$$\begin{aligned}\inf_M \|f - f_M^{NL}\|^2 + T^2 M &\leq \inf_M C M^{-\beta} + T^2 M \\ &\leq \inf_x C x^{-\beta} + T^2(x+1)\end{aligned}$$

differentating this function yields the equation $-\beta C x^{-\beta-1} + T^2 = 0$ and thus

$$\leq C\left(\frac{T^2}{\beta C}\right)^{\beta/(\beta+1)} + T^2\left(\frac{T^2}{\beta C}\right)^{-1/(\beta+1)} + T^2$$
$$x \leq C'\left(T^2\right)^{\beta/(\beta+1)}$$

$2 \Leftrightarrow 3$: The infimum of 2 is exactely the quantity appearing in 3.
$3 \implies 4$, $3 \implies 5$ and (4 and 5) $\implies 3$ are straightforward.
(4 and 5) $\implies 1$ is based on

$$M(T) \leq C(T^2)^{-1/(\beta+1)} \implies (T^2)^{\beta/(\beta+1)} \leq C(M(T))^{-\beta} \quad .$$

$5 \implies 6$:

$$\begin{aligned}\forall T>0, \quad M(T) \leq C(T^2)^{-1/(\beta+1)} &\implies \forall T>0, \quad M(T)^{\beta+1} \leq C^{\beta+1}(T^2)^{-1} \\ &\implies \forall T>0, \quad M(T)^{\beta+1} \leq C^{\beta+1}(T^2) \leq C^{\beta+1}M(T)^{-(\beta+1)}\end{aligned}$$

which yields the results as $M(T)$ takes all the possible values (if we assume that all the coefficients are differents...)
$6 \implies 1$:

$$\begin{aligned}\|f - f_M^{NL}\|^2 &= \sum_{k=M+1}^{+\infty} T(k)^2 &(1)\\ &\leq \sum_{k=M+1}^{+\infty} Ck^{-(\beta+1)} &(2)\\ &\leq CC_\beta M^{-\beta} &(3)\end{aligned}$$

□

With a similar proof than in the linear case, a sufficient condition is given by the reordered Sobolev like condition $\sum_{M=1}^{\infty} M^\beta |T(M)|^2 \leq C$.

2 Basis

For a given function f or a given class of function Θ, the behavior of the coefficients depends obviously of the choice of the basis. We have seen so far that the coefficients should decay rapidly in order to obtain a good approximation rate. In this section, we will look at to families of basis: the ones based on Fourier and the ones based on wavelets.

For each family, we will review the basis, the associated transform and fast transform. We will also see the associated function class with respect to the approximation properties.

2.1 Fourier Basis

2.1.1 Basis

Fourier: most classical basis on $[0, 1]$

FFT algorithm

2.1.2 Issues

Issues with periodization and non localization.

DCT basis that avoids the discontinuity in the periodization

2.1.3 Approximation

Linear approximation class: Sobolev spaces.

Non linear approximation spaces: not well known!

Issues with a step functions.

2.2 Heiseinberg and block based basis

2.2.1 Heisenberg Incertitude Theorem

2.2.2 Block by block basis

2.3 Wavelet Basis

2.3.1 Haar Basis

Explicit Haar basis.

Computation of coefficients $\mathbf{C}^{\alpha}$ and step. Corresponding approximation rate.

2.3.2 Multiresolution and orthogonal wavelet basis

Construction with Fourier theory

Daubechies, Symmlet, Coifflet

2.3.3 Fast Wavelet Transform and Initialization

FWT

First scale initialization.

2.4 Linear and non linear approximation

Besov spaces and Weak Besov spaces.

3 Compression and Approximation Theory

In this section, one will study some very simple compression algorithm to stress the relationship between transform coding compression and non linear approximation.

3.1 Transform coding

To describe a signal f of finite length N, instead of giving its N values, we will use a *transformed* description: its coefficients in an orthogonal basis (b_j).

We thus describe

$$f = \sum_{i=1}^{N} \langle f, b_j \rangle b_j$$

by the set of coefficients $(\langle f, b_j \rangle)$. To code the function f, all we have to code is thus this sequence of coefficients. If we assume that each coefficients requires N_b bits then the total number of coefficients required to code the function f is $N \times N_b$ bits. This number is roughly the same than the one required to give the N values of f.

This yields to a perfect reconstruction of f: a lossless coding. In this setting, the number of bits required can be lowered by using an entropy coding algorithm. These algorithms are based on the information theory and are able to exploit the redundancies in the list of coefficients. These are the algorithm that are used for the ZIP compression for example.

Although this will give, most of the time, a smaller number of bits, this technique is not sufficient to obtain a compression factor greater than 5. To obtain larger compression factor, one has to abandon something: the perfect reconstruction properties.

Indeed, if one is ready to code an approximated version $\widetilde{f}$ of the original signal f, one can further reduce the number of bits required.

3.2 Quantization

The key idea in the transform coding scheme is to code an approximate value of the coefficient instead of the coefficient. This can be done by a simple scalar quantization operator: each coefficient c_n is replaced by a quantized coefficients $Q(c_n)$ where the quantizer maps the real axis into a countable of even finite set of values: it maps a this countable set of bins to a value withing this set.

The simplest quantizer family, the uniform quantizer family, is parameterized by a step size Δ. The quantizer Q_Δ is then the function

$$x \mapsto Q_\delta(x) = \left\lfloor \frac{x + .5\Delta}{\Delta} \right\rfloor \Delta \quad .$$

For any $x \in \mathbb{R}$, $|x - \Delta(x)|^2 \leq \frac{\Delta^2}{4}$. If one only suppose that X is a random variable whose density f satisfy the high resolution hypothesis, $f(x) \simeq f(k\Delta), \forall x \in [(k - .5)\Delta, (k + .5)\Delta]$, one can prove that the uniform quantizer is an optimal strategy to minimize the number of bins for a given error and one obtains $E(|X - Q_\Delta(X)|^2) \simeq \frac{\Delta^2}{12}$.

Applying this scheme to our function f yields

$$\widetilde{f}_\Delta = \sum_{i=1}^{N} Q_\Delta(\langle f, b_j \rangle) b_j$$

and thus

$$\|f - \widetilde{f}_\Delta\|^2 = \sum_{i=1}^{N} |\langle f, b_j \rangle - Q_\Delta(\langle f, b_j \rangle)|^2 \quad .$$

As often the zero bin is very different from this other ones (one hopes that most of the coefficients are small), this error is often rewritten as

$$\|f - \widetilde{f}_\Delta\|^2 = \sum_{Q_\Delta(\langle f, b_j \rangle)=0} |\langle f, b_j \rangle - Q_\Delta(\langle f, b_j \rangle)|^2 + \sum_{Q_\Delta(\langle f, b_j \rangle)\neq 0} |\langle f, b_j \rangle - Q_\Delta(\langle f, b_j \rangle)|^2$$

$$\|f - \widetilde{f}_\Delta\|^2 = \sum_{|\langle f, b_j \rangle| \leq \Delta/2} |\langle f, b_j \rangle|^2 + \sum_{|\langle f, b_j \rangle| > \Delta/2} |\langle f, b_j \rangle - Q_\Delta(\langle f, b_j \rangle)|^2$$

now introducing $f_{\Delta/2}$ the approximation of f made by keeping the coefficients whose absolute values are larger than $\Delta/2$

$$\|f - \widetilde{f}_\Delta\|^2 = \|f - f_{\Delta/2}|^2 + \sum_{|\langle f, b_j \rangle| > \Delta/2} |\langle f, b_j \rangle - Q_\Delta(\langle f, b_j \rangle)|^2$$

Using the simple bound on the quantized error yields

$$\|f - \widetilde{f}_\Delta\|^2 \leq \|f - f_{\Delta/2}|^2 + \left(\frac{\Delta}{2}\right)^2 M_{\Delta/2}$$

which can be proved to be quite tight as under the high resolution hypothesis outside the zero bin

$$\|f - \widetilde{f}_\Delta\|^2 \simeq \|f - f_{\Delta/2}|^2 + \frac{1}{3}\left(\frac{\Delta}{2}\right)^2 M_{\Delta/2}$$

The error $\|f - \widetilde{f}_\Delta\|^2$ is called the distorsion and is denoted D. Its upper bound $\|f - f_{\Delta/2}|^2 + \left(\frac{\Delta}{2}\right)^2 M_{\Delta/2}$ is the quantity that has been studied for approximation and one know thus that its decay is linked to the decay of the coefficients within the basis.

One should know study how to code the set of coefficients.

3.3 Compression without Information Theory

We should first propose a very naïve strategy that would be a first benchmark. Assume that all the coefficients are bounded by $c_{\max}$, the number of bins of size Δ is roughly $\frac{2c_{\max}}{\Delta}$ and thus we need $\log_2 \frac{2c_{\max}}{\Delta}$ bits to specify each coefficient. The resulting rate to code all the coefficients is thus

$$R = N(\log_2 \frac{2c_{\max}}{\Delta}) = N \log_2(2c_{\max}) - \frac{N}{2} \log_2 \Delta^2 \quad .$$

and thus

$$\Delta^2 = 4c_{max}^2 2^{-2R/N}$$

If we look at the average rate $\bar{R}$ this equality becomes

$$\Delta^2 = 4c_{max}^2 2^{-2\bar{R}}$$

If we assume that f is such that it satisfies $\|f - f_M\|^2 \leq CM^{-\beta}$ one can obtain:

$$D(R) \leq C(\Delta^2)^{\beta/(\beta+1)} \leq C' 2^{-2\frac{\beta}{\beta+1}\bar{R}} = C(\Delta^2)^{\beta/(\beta+1)} \leq C' 2^{-2\frac{\beta}{\beta+1}R/N} \quad .$$

One can do much better, once we assume that the proportion of coefficients quantized to zero is large (close to 1). Instead of storing the value of all the coefficients, we will store the number $M_{\Delta/2}$ of significant coefficients, their position and their value. The total number of bits required by this strategy is thus

$$R = \log_2 N + M_{\Delta/2}(\log_2 N + \log_2(2c_{\max}) - \frac{1}{2} \log_2 \Delta^2)$$

As soon as we assume that f is such that it satisfies $\|f - f_M\|^2 \leq CM^{-\beta}$ one can obtain $M_{\Delta/2} \leq (\|f - f_{\Delta/2}|^2)^{-1/\beta}$ and $\Delta^2 \geq (\|f - f_{\Delta/2}|^2)^{(beta+1)/\beta}$ and thus

$$R \leq \log_2 N + D(R)^{-1/\beta} \left(\log_2 N + \log_2(2c_{\max}) - \frac{\beta+1}{2\beta} \log_2 D(R) \right)$$

which yields to

$$D(R) \leq CR^{-\beta} |\log_2 R|^{\beta+1} \quad .$$

3.4 Compression with IT

Shannon...

Peformance closely linked to approximation.

4 Estimation and Approximation Theory

In this section, we focus on the estimation of a finite dimensional signal f from a "noisy" observation $Y = f + \epsilon W$ where W is a standard gaussian white noise and ϵ is a known noise level parameter. This model is the projection on finite dimensional space of the classical gaussian white noise model

$$dY = f(t)dt + \epsilon dW_t$$

and thus we will talk about the regularity of f as the regularity of the underlying continuous function even in the finite dimensional case.

4.1 Oracle estimator and minimax bound

The gaussian white noise model $Y = f + \epsilon W$ can be translated in a sequence model by the decomposition of Y in an orthogonal basis:

$$\langle Y, b_i \rangle = \langle f, b_i \rangle + \epsilon \langle W, b_i \rangle \quad .$$

Using the properties of a white noise, one can verify that the sequence $(\langle W, b_i \rangle)$ is standard gaussian iid.

Oracle estimator / Minimax lower bound for orthosymmetric bodies?

4.2 Thresholding estimator and model selection

Thresholding estimator.

Let $\mathcal{D}$ be a dictionnary of orthogonal basis $\mathcal{B}$, we call models $\mathcal{M}_{\gamma \in \Gamma}$ the set of all subspaces spanned by generators $(e_i)_{i \in I}$ belonging to a single basis $\mathcal{B}$ of $\mathcal{D}$.

To build a good estimate F of f from Y, we need then to build a good dictionnary $\mathcal{D}$ and find a feasable way to select a good model $\mathcal{M}$.

A good dictionnary $\mathcal{D}$ may obey several constraints

- $\mathcal{D}$ must be rich enough to ensure that any $\mathbf{C}^\alpha$ geometrically regular function f may be well approximated in at least one basis $\mathcal{B}$ of $\mathcal{D}$.
- The number of generators of all basis of $\mathcal{D}$ must be finite and not too high to allow a feasible algorithm to define the estimatate F and to prevent overfitting problems.

The last point is crucial. The choice of the basis $\mathcal{B}$ depends on data Y and then on $P_{\mathcal{L}} dW$. This dependency to realization of noise and the large number of models could lead to overfitting problems. To overcome this difficulty Barron, Birgé and Massart[1] propose to select the model $\mathcal{M}_F$ that minimizes a penalized criterion

$$\widehat{\mathcal{M}} = \operatorname*{argmin}_{\mathcal{M}_\gamma, \gamma \in \Gamma} \|P_{\mathcal{L}} Y - P_{\mathcal{M}_\gamma} Y\|^2 + \text{pen}(\gamma)$$

with a suitable choice for $\text{pen}(\gamma)$. $\text{pen}(\gamma)$ penalizes the complexity of the model to minimize the dimension of the model and then the energy of $P_{\mathcal{M}_\gamma}(dW)$.

Indeed one can prove

Theorem 1. *Let $\{\mathcal{M}_\gamma\}_{\gamma\in\Gamma}$ be a collection of subspaces generated by generators taken in a finite set $\mathcal{C}$ of cardinal κ, let $\mathcal{L}$ a finite dimensional subspace such that $\mathcal{M}_\gamma \subset \mathcal{L}, \forall \gamma \in \Gamma$, let $\lambda \geq \sqrt{32 + \frac{8}{\log \kappa}}$. For any $f \in L^2$ and any noise level ε the estimate F defined from $\mathrm{d}Y_t = f(t)\mathrm{d}t + \varepsilon \mathrm{d}W_t$ as the projection $F = P_{\widehat{\mathcal{M}}} Y$ of the noise on the best model*

$$\widehat{\mathcal{M}} = \underset{\mathcal{M}_\gamma, \gamma\in\Gamma}{argmin} \|P_{\mathcal{L}} Y - P_{\mathcal{M}_\gamma} Y\|^2 + \lambda^2 (\log \kappa) \varepsilon^2 M_\gamma \quad .$$

satisfies

$$E(\|f - F\|^2) \leq 4 \left(\min_{\mathcal{M}_\gamma, \gamma\in\Gamma} \|f - P_{\mathcal{M}_\gamma} f\|^2 + \lambda^2 (\log \kappa) \varepsilon^2 M_\gamma \right) + \frac{32}{\kappa} \varepsilon^2$$

where M_F is the dimension of the subspace on which F lives for a constant C that depends only on λ.

We propose here a simple proof, inspired by a sketch of Barron, Birgé, Massart, based only on a concentration lemma for the norm of the projection of the noise over all subspaces deduced from an inequality for gaussian process obtained by Tsirelson, Ibragimov and Sudakov[42]. A better lower bound on λ can be obtained[34] but at the price of some technicalities.

Concentration inequalities are at the core of all the selection model estimators. Essentially, the penalty should dominate the random fluctuation of the minimized quantity. The key lemma, Lemma 1, uses a concentration inequality for gaussian variable to ensure, with high probability, that the noise energy is small simultaneously in all the subspaces $\mathcal{M}_\gamma$.

Lemma 1. *For all $u \geq 0$, with a probability greater than $1 - 2/\kappa e^{-u}$,*

$$\forall \gamma \in \Gamma, \|P_{\mathcal{M}_\gamma} \overline{W}\| \leq \sqrt{M_\gamma} + \sqrt{4 \log \kappa M_\gamma + 2u}$$

where $\overline{W} = P_{\mathcal{L}} \mathrm{d}W_t$ is the orthogonal projection of $\mathrm{d}W_t$ on the space $\mathcal{L}$.

Proof of Lemma 1. The key ingredient of this proof is a concentration inequality. Tsirelson's Lemma proves that for any 1-lipschitz function $\phi : \mathbb{R}^n \to \mathbb{R}$ ($|\phi(x) - \phi(y)| \leq \|x - y\|$) if X is a gaussian white noise with variance σ in $\mathbb{R}^n$ then

$$\mathbb{P}\left(\phi(X) \geq E\left(\phi(X)\right) + \sigma t\right) \leq e^{-t^2/2} \quad .$$

Let $\mathcal{L}$ be a subspace of finite dimension and $\mathcal{M}$ be a subspace of dimension M, $\phi_{\mathcal{M}}$ is defined as a function of $\mathcal{L}$ into $\mathbb{R}$ by

$$\phi_{\mathcal{M}}(f) = \|P_{\mathcal{M}}(f)\|$$

which is 1-Lipschitz. Tsirelson's Lemma applies for $t = \sqrt{4 \log \kappa M + 2u}$ yielding for $\overline{W} = P_{\mathcal{L}} dW_t$ which is white noise on $\mathcal{L}$

$$\mathbb{P}\left(\|P_{\mathcal{M}} \overline{W}\| \geq E(\|P_{\mathcal{M}} \overline{W}\|) + \sqrt{4 \log \kappa M + 2u} \right) \leq \kappa^{-2M} e^{-u} \quad .$$

Now as $E(\|P_{\mathcal{M}}\overline{W}\|) \leq (E(\|P_{\mathcal{M}}\overline{W}\|^2))^{1/2} = \sqrt{M}$, one derives

$$\mathbb{P}\left(\|P_{\mathcal{M}}\overline{W}\| \geq \sqrt{M} + \sqrt{4\log\kappa M + 2u}\right) \leq \kappa^{-2M} e^{-u} \quad .$$

Noticing now that

$$\begin{aligned}
\mathbb{P}\left(\exists\gamma\in\Gamma, \|P_{\mathcal{M}_\gamma}\overline{W}\| \geq \sqrt{M_\gamma} + \sqrt{4\log\kappa M_\gamma + 2u}\right) &\leq \sum_{\gamma\in\Gamma} \mathbb{P}\left(\|P_{\mathcal{M}_\gamma}\overline{W}\| \geq \sqrt{M_\gamma} + \sqrt{4\log\kappa M_\gamma + 2u}\right) \\
&\leq \sum_{\gamma\in\Gamma} \kappa^{-2M_\gamma} e^{-u} \\
&\leq \sum_{n=1}^{\kappa} \binom{n}{\kappa} \kappa^{-2n} e^{-u} \leq \sum_{n=1}^{\kappa} \kappa^{-n} e^{-u} \\
&\leq \frac{\kappa^{-1}}{1-\kappa^{-1}} e^{-u} \\
\mathbb{P}\left(\exists\gamma\in\Gamma, \|P_{\mathcal{M}_\gamma}\overline{W}\| \geq \sqrt{M_\gamma} + \sqrt{4\log\kappa M_\gamma + 2u}\right) &\leq \frac{2}{\kappa} e^{-u}
\end{aligned}$$

□

The proof of Theorem 1 given in Appendix **??** defines the best oracle model $\mathcal{M}_O$ as the minimize of the deterministic quantity

$$\min_{\gamma\in\Gamma} \|f - P_{\mathcal{M}_\gamma} f\|^2 + \lambda^2 \log\kappa\varepsilon^2 M_\gamma$$

and let $f_O = P_{\mathcal{M}_O} f$ and $M_O = \dim(\mathcal{M}_O)$. Lemma 1 is used to obtain for all $u > 0$

$$P(\|f - F\|^2 - 4(\|f - f_O\|^2 + \varepsilon^2\lambda^2 \log\kappa M_O) \geq 32\varepsilon^2 u) \leq \frac{2}{\kappa} e^{-u}$$

which implies the bound of Theorem 1 by integration over u.

Corollary 1. *For the choice of the precision $\nu = \varepsilon$ in a wavelet basis, Theorem 1 implies Theorem* **??** *up to the constants.*

Proof. For this choice $N^{-1/2} = \nu$, $\mathcal{L} = V_\nu$, $\mathcal{M}_\gamma$ is the set of all subspaces spanned by some wavelet of scale larger than ν and $\kappa \leq C\varepsilon^{-(p+5)}$.

The minimization of

$$\widehat{\mathcal{M}} = \operatorname*{argmin}_{\mathcal{M}_\gamma, \gamma\in\Gamma} \|P_{\mathcal{L}} Y - P_{\mathcal{M}_\gamma} Y\|^2 + \lambda^2 \log\kappa\varepsilon^2 M_\gamma$$

selects the coefficients that are above a threshold $T = \lambda\sqrt{\log\kappa}\varepsilon$ which is up to the constant the threshold proposed by Donoho and Johnstone. The estimators in both Theorems are thus the same for a suitable choice of λ.

Now

$$
\begin{aligned}
E(\|f-F\|^2) &\leq C \min_{\gamma\in\Gamma} \|f - P_{\mathcal{M}_\gamma} f\|^2 + \lambda^2 \log N \varepsilon^2 M_\gamma + \varepsilon^2 \\
&\leq C\lambda^2 \log N \min_{\gamma\in\Gamma} \|f - P_{\mathcal{M}_\gamma} f\|^2 + \varepsilon^2 M_\gamma + \varepsilon^2 \\
E(\|f-F\|^2) &\leq C\lambda^2 \log N (E(\|f-F_O\|^2) + \varepsilon^2)
\end{aligned}
$$

and thus the bound of Theorem **??** is obtained. □

The deterministic quantity

$$
\min_{\gamma\in\Gamma} \|f - P_{\mathcal{M}_\gamma} f\|^2 + \lambda^2 \log \kappa \varepsilon^2 M_\gamma + \frac{8}{\kappa}\varepsilon^2
$$

is thus the analog of the oracle risk of the thresholding estimator and therefore will also be called oracle risk.

This oracle risk

$$
\min_{\gamma\in\Gamma} \|f - P_{\mathcal{M}_\gamma} f\|^2 + \lambda^2 \log \kappa \varepsilon^2 M_\gamma + \varepsilon^2
$$

depends on the precision ν through the collection $\{\mathcal{M}_\gamma\}_{\gamma\in\Gamma}$. This precision should depend on the noise level ε. Indeed, on the one hand, the number κ of generators of the models should be controlled so that $\log \kappa$ remains small comparing to ε^2. On the other hand, the models should be rich enough to guaranty that the minimum of the oracle risk is small. This leads to a trade-off which is satisfied for example in most case by letting $\nu = \varepsilon$ for both the wavelet and the curvelets.

Proof. Let $\bar{Y} = P_{\mathcal{L}} Y$, $\bar{f} = P_{\mathcal{L}} f$. Recall that

$$
F = \operatorname*{argmin}_{\substack{\tilde{f} = P_{\mathcal{M}_\gamma} \bar{f} \\ \gamma\in\Gamma}} \|\bar{Y} - \tilde{f}\|^2 + \lambda^2 \log \kappa \varepsilon^2 \dim(\mathcal{M}_\gamma) \ .
$$

Let $\widehat{\mathcal{M}}$ the corresponding subspace and M_F its dimension

Define

$$
f_O = \operatorname*{argmin}_{\substack{\tilde{f} = P_{\mathcal{M}_\gamma} \bar{f} \\ \gamma\in\Gamma}} \|\bar{f} - \tilde{f}\|^2 + \lambda^2 \log \kappa \varepsilon^2 \dim(\mathcal{M}_\gamma) \ .
$$

Let $\mathcal{M}_O$ the corresponding subspace and M_O its dimension.

By construction,

$$
\|\bar{Y} - F\|^2 + \lambda^2 \log \kappa \varepsilon^2 M_F \leq \|\bar{Y} - f_O\|^2 + \lambda^2 \log \kappa \varepsilon^2 M_O
$$

using $\|\bar{Y} - F\|^2 = \|\bar{Y} - \bar{f}\|^2 + \|\bar{f} - F\|^2 + 2\langle \bar{Y} - \bar{f}, \bar{f} - F\rangle$ and a similar equality for $\|\bar{Y} - f_O\|^2$, one obtains

$$
\begin{aligned}
\|\bar{f} - F\|^2 + \lambda^2 \log \kappa \varepsilon^2 M_F \leq \|\bar{f} - f_O\|^2 &+ \lambda^2 \log \kappa \varepsilon^2 M_O \\
&+ 2\langle \bar{Y} - \bar{f}, F - f_O\rangle
\end{aligned}
$$

One should now concentrate on the bound on the scalar product :

$$
\begin{aligned}
|2\langle \bar{Y} - \bar{f}, F - f_O\rangle| &= |2\langle \varepsilon P_{\widehat{\mathcal{M}} \cup \mathcal{M}_O} \overline{W}, F - f_O\rangle| \\
&\leq 2\varepsilon \|P_{\widehat{\mathcal{M}} \cup \mathcal{M}_O} \overline{W}\| (\|\bar{f} - F\| + \|\bar{f} - f_O\|)
\end{aligned}
$$

Using Lemma 1, with a probability greater than $1 - \frac{2}{\kappa} e^{-u}$

$$
\leq 2\varepsilon (\sqrt{M_O + M_F} + \sqrt{4 \log \kappa (M_O + M_F) + 2u}) (\|\bar{f} - F\| + \|\bar{f} - f_O\|)
$$

and using $2xy \leq \beta^2 x^2 + \beta^{-2} y^2$ successively with $\beta = \alpha$ and $\beta = 1$

$$
|2\langle \bar{Y} - \bar{f}, F - f_O\rangle| \leq (\alpha^2 2 (\|\bar{f} - f_O\|^2 + \|\bar{f} - F\|^2) + \alpha^{-2} 2\varepsilon^2 (M_O + M_F + 4 \log \kappa (M_O + M_F) + 2u)) \quad .
$$

This leads to

$$
\begin{aligned}
(1 - 2\alpha^2) \|f - F\|^2 \leq (1 + 2\alpha^2) &\|f - f_O\|^2 \\
&+ \varepsilon^2 (\lambda^2 \log \kappa + 2\alpha^{-2} (1 + 4 \log \kappa)) M_O \\
&+ \varepsilon^2 (2\alpha^{-2} (1 + 4 \log \kappa) - \lambda^2 \log \kappa) M_F \\
&+ 4\varepsilon^2 \alpha^{-2} u
\end{aligned}
$$

Choosing $\alpha = \frac{1}{2}$ gives

$$
\begin{aligned}
\frac{1}{2} \|f - F\|^2 \leq \frac{3}{2} &\|f - f_O\|^2 \\
&+ \varepsilon^2 (\lambda^2 \log \kappa + 8(1 + 4 \log \kappa)) M_O \\
&+ \varepsilon^2 (8(1 + 4 \log \kappa) - \lambda^2 \log \kappa) M_F \\
&+ 16\varepsilon^2 u
\end{aligned}
$$

So that if $\lambda^2 \geq 32 + \frac{8}{\log \kappa}$

$$
\begin{aligned}
\|f - F\|^2 \leq 3 &\|f - f_O\|^2 \\
&+ 4\varepsilon^2 \lambda^2 \log \kappa M_O \\
&+ 32\varepsilon^2 u
\end{aligned}
$$

which implies

$$
\|f - F\|^2 \leq 4 (\|f - f_O\|^2 + \varepsilon^2 \lambda^2 \log \kappa M_O) + 32\varepsilon^2 u
$$

where this results holds with probability greater than $1 - \frac{2}{\kappa} e^{-u}$.

Recalling that this is valid for all $u \geq 0$, one has

$$
P(\|f - F\|^2 - 4(\|f - f_O\|^2 + \varepsilon^2 \lambda^2 \log \kappa M_O) \geq 32\varepsilon^2 u) \leq \frac{2}{\kappa} e^{-u}
$$

which implies by integration over u

$$E(\|f-F\|^2 - 4(\|f-f_O\|^2 + \varepsilon^2\lambda^2 \log \kappa M_O)) \leq 32\varepsilon^2 \frac{2}{\kappa}$$

that is the bound of Theorem 1

$$E(\|f-F\|^2) \leq 4(\|f-f_O\|^2 + \varepsilon^2\lambda^2 \log \kappa M_O) + 32\varepsilon^2 \frac{1}{\kappa} \quad .$$

□

5 Geometrical Representation

References

[1] A. Barron, L. Birge, and P. Massart. Risk bounds for model selection via penalization. *Probab. Th. Rel. Fields*, 113:301–413, 1999.

[2] L. Birgé and P. Massart. From model selection to adaptive estimation. In D. Pollard, E. Torgersen, and G. L. Yang, editors, *A Festschrift for Lucien Le Cam*, pages 55–87. New York, 1995.

[3] E. Candès and D. Donoho. Curvelets: A surprisingly effective nonadaptive representation of objects with edges. In L. L. Schumaker, A. Cohen, and C. Rabut, editors, *Curves and Surfaces fitting*. Vanderbilt University Press, 1999.

[4] E.J. Candès, D.L. Donoho, and J.L Stark. The curvelet transform for image denoising. *IEEE Transactions on Image Processing*, (11):670–684, 2000.

[5] E. Candès. Modern statistical estimation via oracle inequalities. *Acta Numerica*, 2006.

[6] E.J. Candès and D.L. Donoho. A surprinsingly effective nonadaptive representation for objects with edges. *Curves and Surfaces*, 1999.

[7] A. Cohen, I. Daubechies, and P. Vial. Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harm. Anal.*, 1:54–81, 1993.

[8] A. Cohen, R. DeVore, G. Kerkyacharian, and D. Picard. Maximal spaces with given rate of convergence for thresholding algorithms. *Appl. Comput. Harmon. Anal.*, 11(2):167–191, 2001.

[9] A. Cohen, R. deVore, P. Petrushev, and H. Xu. Non linear approximation and the space $BV(\mathbb{R}^2)$. *Amer. J. Math*, 121:587–628, 1999.

[10] A. Cohen and B. Matei. Compact representations of images by edge adapted multiscale transforms. Thessalonique, October 2001.

[11] R. Coifman and D. Donoho. Translation-invariant denoising. In *Wavelet and Statistics*, Lecture Notes in Statistics. Springer Verlag, 1995.

[12] JPEG200 Committee. Jpeg 2000 image coding system, part1. *ISO/IEC JTC 1/SC 29/WG 1 N2678, 19*, July 2002.

[13] I. Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1992.

[14] I. Daubechies. Orthonormal bases of compactly supported wavelets. *Commun. on Pure and Appl. Math.*, 36(5):961–1005, September 1998.

[15] R. DeVore. Nonlinear approximation. *Acta. Numer.*, 7:51–150, 1998.

[16] R deVore, B. Jawerth, and B. Lucier. Image compression through wavelet transform coding. *IEEE Trans. Info. Theory*, 38(3):719–746, March 1992.

[17] M Do. *Directional Multiresolution Image Representations*. PhD thesis, Department of Communication Systems, Swiss Federal Institute of Technology Lausanne, November 2001.

[18] M. Do and M. Vetterli. Curvelets and filter banks. *IEEE Trans. Image Proc.*, 2001.

[19] D. Donoho. Wedgelets: Nearly-minimax estimation of edges. *Ann. Statist*, 27:353–382, 1999.

[20] D. Donoho and I. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, December 1994.

[21] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Universal near minimaxity of wavelet shrinkage. 1995.

[22] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: Asymptopia? (with discussion). *J. of the Royal Stat. Soc., Series B*, 57:301–369, 1995.

[23] D.L. Donoho. Unconditional bases are optimal bases for data compression and statistical estimation. *Applied and Computational Harmonic Analysis*, 1(1):100–115, 1993.

[24] D.L. Donoho and I.M. Johnstone. Ideal denoising in an orthonormal basis chosen from a library of bases. *Comptes Rendus de l'Académie des Sciences*, Serie 1(319):1317–1322, 1994.

[25] F. Falzon and S. Mallat. Analysis of low bit rate image transform coding. *IEEE Transaction on Signal Processing*, January 1998.

[26] JPEG 2000, verification model 5.0, ISO/IEC JTC1/SC29/WG1 N1429, 2002.

[27] N. Kingsbury. Complex wavelets for shift invariant analysis and filtering of signals. *Journal of Appl. and Comput. Harmonic Analysis*, 10:234–253, 2001.

[28] E.D. Kolaczyk and R.D. Nowak. Multiscale likelihood analysis and complexity penalized estimation. *Annals of Statistics*, 32:500–527, 2004.

[29] A. Korostelev and A. Tsybakov. *Minimax Theory of Image Reconstruction*, volume 82 of *Lecture Notes in Statistics*. Springer, 1993.

[30] E. Le Pennec and S. Mallat. Sparse Geometrical Image Approximation with Bandelets. *IEEE Transaction on Image Processing*, 14(4):423–438, 2004.

[31] E. Le Pennec and S. Mallat. Bandlet image approximation and compression. *SIAM Multiscale Modeling and Simulation*, 4(3):992–1039, 2005.

[32] S. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. 11(7):674–693, July 1989.

[33] S. Mallat. *A wavelet tour of signal processing*. Academic Press, 2nd edition edition, 1998.

[34] P. Massart. *Concentration Inequalities and Model Selection (Saint Flour Notes)*. Springer, 2003.

[35] G. Peyré and S. Mallat. Orthogonal bandlets bases for geometric images approximation. Technical report, CMAP, 2007.

[36] J. Portilla, V. Strela, M. Wainwright, and E. P. Simoncelli. Image denoising using a scale mixture of Gaussians in the wavelet domain. *IEEE Trans Image Processing*, 12(11):1338–1351, November 2003.

[37] K. Rao and P. Yip. *Discret Cosine transform: Algorithms, Advantages, Applications*. Academic Press, 1990.

[38] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[39] J. Shapiro. An embedded hierarchical image coder using zerotrees of wavelet coefficients. In *Proc. IEEE Data Compression Conference*, pages 214–223, Snowbird, UT, March 1993.

[40] J. Starck, E. Candès, and D. Donoho. The curvelet transform for image denoising. (soumis à IEEE Transactions on Signal Processing).

[41] D. Taubman. EBCOT (embedded block coding with optimized truncation): A complete reference. ISO/IEC JTC1/SC29/WG1 N983, September 1998.

[42] B.S. Tsirelson, I.A. Ibragimov, and V.N. Sudakov. Norms of gaussian sample functions. In *Lecture Notes in Mathematics*, volume 550, pages 20–41. Springer, 1976.